What choice(s) do we have? Processing and contextual constraints on syntactic variation across the globe

Melanie Röthlisberger, Jason Grafmiller, Benedikt Heller & Benedikt Szmrecsanyi

KU Leuven
Quantitative Lexicology and Variational Linguistics

Introduction

Introduction

- "Exploring probabilistic grammar(s) in varieties of English around the world" (5-year project, 2013–2018; PI: Benedikt Szmrecsanyi)
- synthesize disjoint lines of scholarship – research on World Englishes & Probabilistic Grammar – into one unifying project with a coherent empirical and theoretical focus
- main goal: understand the plasticity of probabilistic knowledge of grammar, on the part of language users with diverse regional and cultural backgrounds

Introduction

- "Exploring probabilistic grammar(s) in varieties of English around the world" (5-year project, 2013–2018; PI: Benedikt Szmrecsanyi)
- synthesize disjoint lines of scholarship – research on World Englishes & Probabilistic Grammar – into one unifying project with a coherent empirical and theoretical focus
- main goal: understand the plasticity of probabilistic knowledge of grammar, on the part of language users with diverse regional and cultural backgrounds
- today: variation across three syntactic alternations in 9 international varieties of English

Today

1. Introduction
2. Method & Data
3. Findings
4. Concluding remarks

The "English World-Wide Paradigm"

- wide range of postcolonial varieties (e.g. Hong Kong E), inner circle varieties (e.g. British E), shift varieties (e.g. Irish E),...
- topics: scope, limits, parameters of variation; extent to which structural make-up of varieties of E can be predicted by communicative needs of colonizers/colonized
  (e.g. Kachru 1992; Schneider 2007; Mesthrie and Bhatt 2008)
- shortcoming: an often primarily descriptive interest in the variable presence/absence of features, or in usage frequencies of features

The Probabilistic Grammar framework

- rely on the variation-centered, usage- and experience-based probabilistic grammar framework developed by Joan Bresnan and collaborators

  (e.g. Bresnan et al. 2007; Bresnan and Ford 2010; Wolk et al. 2013)

  1. syntactic variation – and change – is **subtle, gradient & probabilistic** rather than categorical in nature

     (Labov 1982; Bresnan and Hay 2008)
  2. linguistic knowledge includes **knowledge of probabilities**, and speakers have powerful predictive capacities

     (Gahl and Garnsey 2004; Gahl and Yu 2006)

The Probabilistic Grammar framework

"A probabilistic, usage-based approach to grammar is able to account for [...] variation by assuming that different communities differ in the types and frequencies of the constructions that they are exposed to. However, **a probabilistic approach also predicts that variation across space and time should exist in less obvious ways** - even affecting the **subtle probabilistic choices** that are made between two variants which are equally acceptable for that dialect. [...], we expect to observe syntactic differences in time and space which are reflected [...] in extremely subtle factors such as the **relative probabilistic weights of conditioning factors** [...]."

Bresnan and Hay 2008: 246

Some research questions

- scope and limits of variation – do the varieties of English we study here share a core probabilistic grammar?
- dialect typology – does variety type (e.g. inner versus outer circle) predict probabilistic similarity between varieties of English?
- variation phenomena – do the alternations under study differ in terms of their probabilistic sensitivity to variety effects?

Method & Data

A methodological sketch of the project

1. tap into corpus data to explore 3 syntactic alternations across 9 varieties

A methodological sketch of the project

1. tap into corpus data to explore 3 syntactic alternations across 9 varieties
2. use the variationist method (Labov 1982) to create richly annotated corpus-derived datasets . . .

A methodological sketch of the project

1. tap into corpus data to explore 3 syntactic alternations across 9 varieties
2. use the variationist method (Labov 1982) to create richly annotated corpus-derived datasets . . .
3. . . . to study the interplay of probabilistic factors constraining the alternations; check for significant differences between varieties
4. last stage of the project: conduct supplementary rating-task experiments

Varieties of English

- British E, Canadian E, Indian E, Singapore E, Irish E, New Zealand E, Hong Kong E, Jamaican E, Philippines E
- corpus database: 1.5m words of running text per variety, covering spoken/written English (ICE), and (eventually) web-based language (GloWbE)

genitive alternation

(1) a. [The Senator]$_{possessor}$'s [brother]$_{possessum}$
(the *s*-genitive)
b. [The brother]$_{possessum}$ of [the Senator]$_{possessor}$
(the *of*-genitive)

- variable context: identified *'s* and *of* occurrences; manually excluded e.g. partitive genitives and pronominal genitives; N=10,592
- investigator: Benedikt Heller

dative alternation

(2) a. We sent [the president]$_{recipient}$ [a letter]$_{theme}$
(the ditransitive dative)
b. We sent [a letter]$_{theme}$ to [the president]$_{recipient}$
(the prepositional dative)

- variable context: used a list of dative verbs to identify occurrences; manually excluded e.g. passivized verbs, elliptic structures, etc.; N=8,549
- investigator: Melanie Röthlisberger

particle placement

(3)
a. The president looked$_{verb}$ [the word]$_{NP}$ up$_{particle}$ (V-DO-P)
b. The president looked$_{verb}$ up$_{particle}$ [the word]$_{NP}$ (V-P-DO)

- variable context: transitive particle verbs involving one of the following 10 particles: *around*, *away*, *back*, *down*, *in*, *off*, *out*, *over*, *on*, *up*; manually excluded e.g. passive sentences and sentences with extracted direct objects; N=8,072
- investigator: Jason Grafmiller

Annotation

- predictors across alternations:
  constituent length (➪ end weight); animacy; constituent givenness; thematicity; TTR; overall frequency of head nouns; genre; variety
- alternation specific predictors:
  e.g. presence of directional PPs after particle verb constructions; final sibilancy of genitive possessors; definiteness (of direct object); NP expression type (common noun, pronoun, . . .), idiomaticity of verb-particle, . . .

regression analysis

- logistic regression probes the probabilistic conditioning of linguistic choice-making
- based on annotated linguistic observations, investigates the influence of (more than 1) constraint on a binary outcome (e.g. ditransitive vs prepositional dative)
- checks whether predictors have significant effect and estimates effect size and direction
- `glmer()`, `lrm()`, `glm()` functions in Rs `lme4` package

(Bates, Maechler, and Bolker Bates et al.; Harrell 2001)

Findings

Some first findings

- 3 alternations $\times$ nine varieties
- ICE data only; comparatively simple annotation; logistic regression modeling
- key findings:
  - varieties do share a core probabilistic grammar
  - indigenization at various degrees of subtlety, depending on abstractness of patterns
  - cf. Szmrecsanyi et al. 2016

Do the varieties of English we study share a core probabilistic grammar?

- yes, in the sense that there clearly are variety-independent, qualitative generalizations
- the **effect directions** of factors are stable across varieties of English – but interesting differences with regard to **effect size**

Probabilistic differences in the genitive alternation

Compared to GB...

- effect of animacy of possessor (HKE, NZE, PhiE)
- effect of final sibilancy of possessor (HKE, IndE)
- effect of possessum length (CanE, HKE, IrE, PhiE, SinE)

Probabilistic differences in the genitive alternation

(*the [Senator]'s brother* vs *the brother of the [Senator]*)

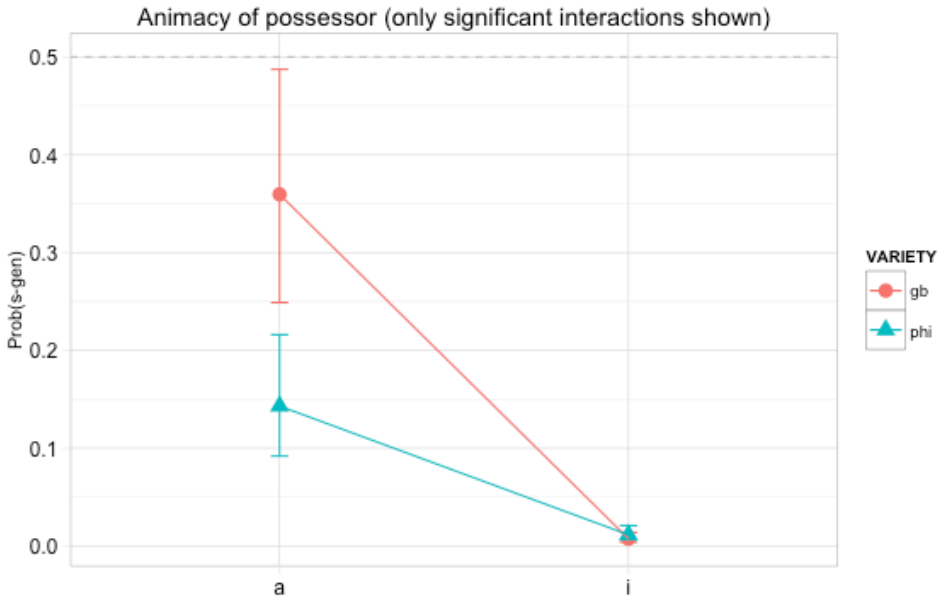


Figure: Predicted probability of the *s*-genitive obtained from mixed-effect model (with 95% confidence intervals)

Probabilistic differences in the dative alternation

Compared to all varieties...

- effect of length (IndE, JamE)
- effect of recipient pronominality (IndE, CanE, JamE)
- effect of register/style (NZE, IrE, JamE, HKE)
- effect of theme concreteness (CanE)

Probabilistic differences in the dative alternation

(*send [him] a letter* vs. *send a letter to [the Senator]*)

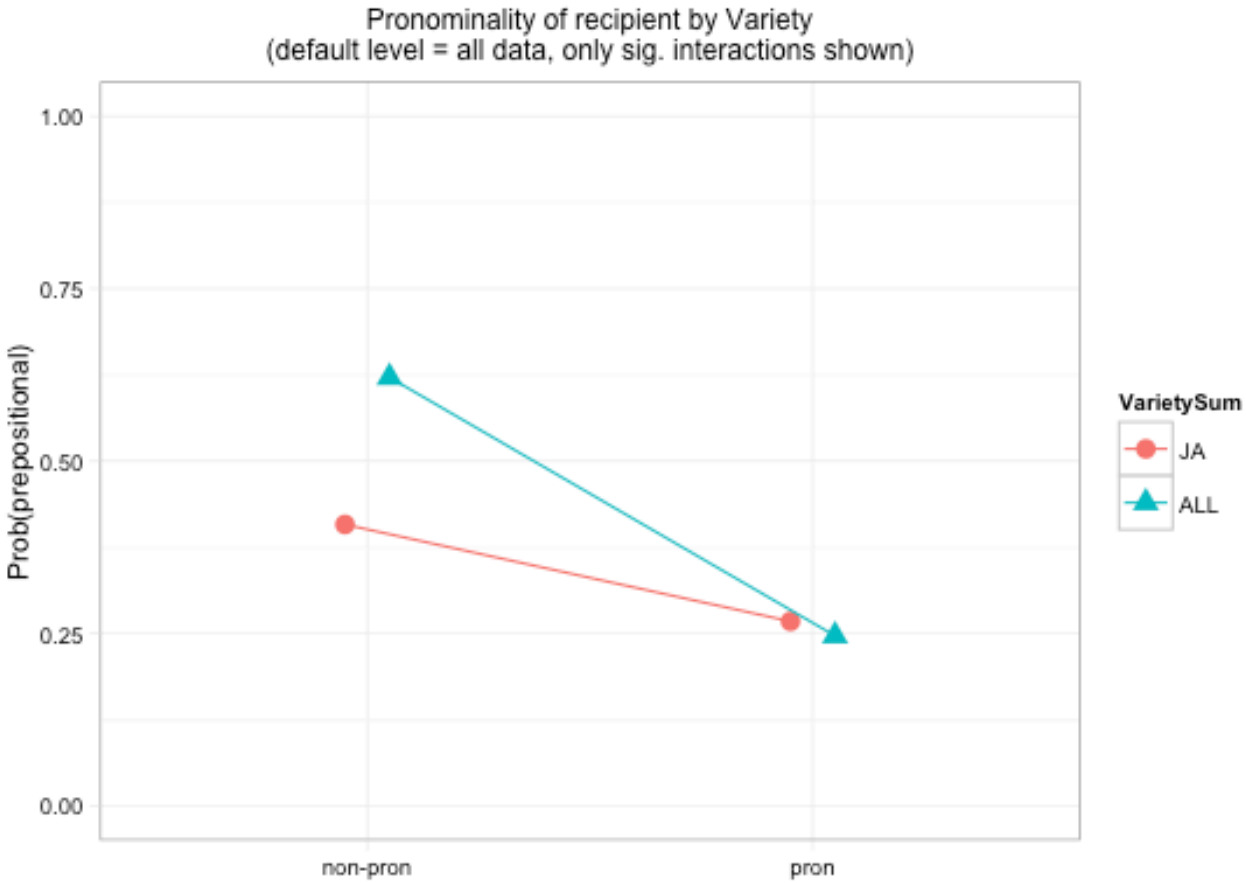


Figure: Predicted probability of the prepositional dative obtained from mixed-effect modeling

Probabilistic differences in particle placement

Compared to GB...

- effect of length (PhiE)
- effect of idiomaticity (NZE)
- effect of concreteness (PhiE, NZE)
- effect of givenness (SinE, IndE)
- effect of presence of post-modifying PP (IndE)

Probabilistic differences in particle placement

(*put [the book] back [on the table] vs put back [the book] [on the table]*)

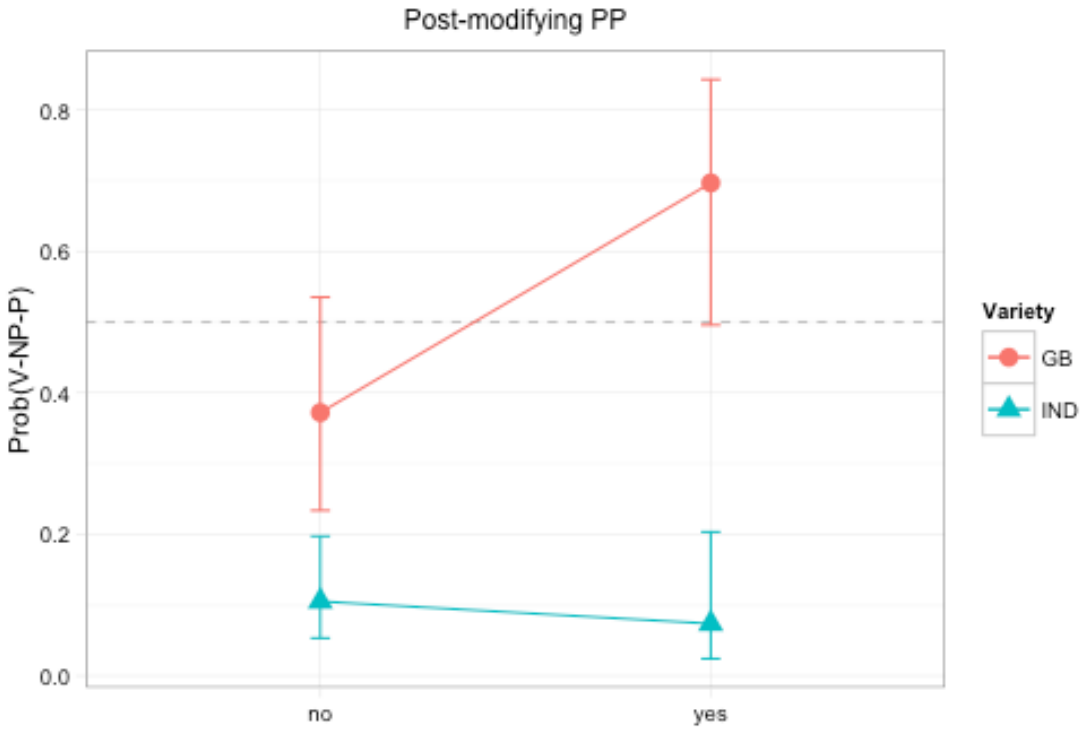


Figure: Predicted probabilities for the split order in Indian English obtained from the mixed-effect modeling (with 95% confidence intervals)

Do we find a split between inner and outer circle varieties of English?

- indications: IndE, SinE, NZE, HKE, PhilE, JamE
- inconclusive patterns

Do the alternations under study differ in terms of their probabilistic sensitivity to variety effects?

- amenability to "probabilistic indigenization":
  - most amenable: particle placement
  - less amenable: genitive alternation, dative alternation
- Schneider (2003: 249): lexico-grammar is a prime target of early-stage indigenization
- tentative generalization: the more tightly associated a given syntactic alternation is with concrete instantiations involving specific lexical items the more likely it is to exhibit cross-varietal indigenization effects

Concluding remarks

What's new

- crossroads of research on English as a World Language, usage-based theoretical linguistics, variationist linguistics and cognitive sociolinguistics

What's new

- crossroads of research on English as a World Language, usage-based theoretical linguistics, variationist linguistics and cognitive sociolinguistics
- interest in scope and limits of variation in a large-scale comparative perspective

What's new

- crossroads of research on English as a World Language, usage-based theoretical linguistics, variationist linguistics and cognitive sociolinguistics
- interest in scope and limits of variation in a large-scale comparative perspective
- assume that language users implicitly learn the probabilistic effects of constraints on variation by constantly (re-)assessing input of spoken and written discourses throughout their lifetimes

Team members

Benedikt Szmrecsanyi
Principle investigator

Jason Grafmiller
Ph.D., 2013, Stanford University
particle placement

Benedikt Heller
MA, 2013, University of Giessen
the genitive alternation

Melanie Röthlisberger
MA, 2011, University of Zurich
the dative alternation

Thank you!

melanie.rothlisberger@kuleuven.be

http://wwwling.arts.kuleuven.be/qlvl/ProbGrammarEnglish.html

This presentation is based upon work supported by an Odysseus grant of the Research Foundation Flanders (FWO) (grant no. G.0C59.13N).

References I

Bates, D. M., M. Maechler, and B. Bolker. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7.

Bresnan, J., A. Cueni, T. Nikitina, and H. Baayen (2007). Predicting the Dative Alternation. pp. 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, J. and M. Ford (2010). Predicting Syntax: Processing dative constructions in American and Australian Varieties of English. *Language 86*(1), 168–213.

Bresnan, J. and J. Hay (2008, February). Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua 118*(2), 245–259.

Gahl, S. and S. Garnsey (2004). Knowledge of Grammar, Knowledge of Usage: Syntactic Probabilities Affect Pronunciation Variation. *Language 80*, 748–775.

Gahl, S. and A. C. Yu (2006). *Special theme issue: Exemplar-based models in linguistics*. The linguistic review. Mouton de Gruyter.

Harrell, F. E. J. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Kachru, B. B. (Ed.) (1992). *The Other tongue: English across cultures* (2nd ed ed.). English in the global context. Urbana: University of Illinois Press.

Labov, W. (1982). Building on empirical foundations. In W. Lehmann and Y. Malkiel (Eds.), *Perspectives on Historical Linguistics*, pp. 17–92. Amsterdam, Philadelphia: Benjamins.

References II

Mesthrie, R. and R. M. Bhatt (2008). *World Englishes: The Study of New Linguistic Varieties*. Cambridge: Cambridge University Press.

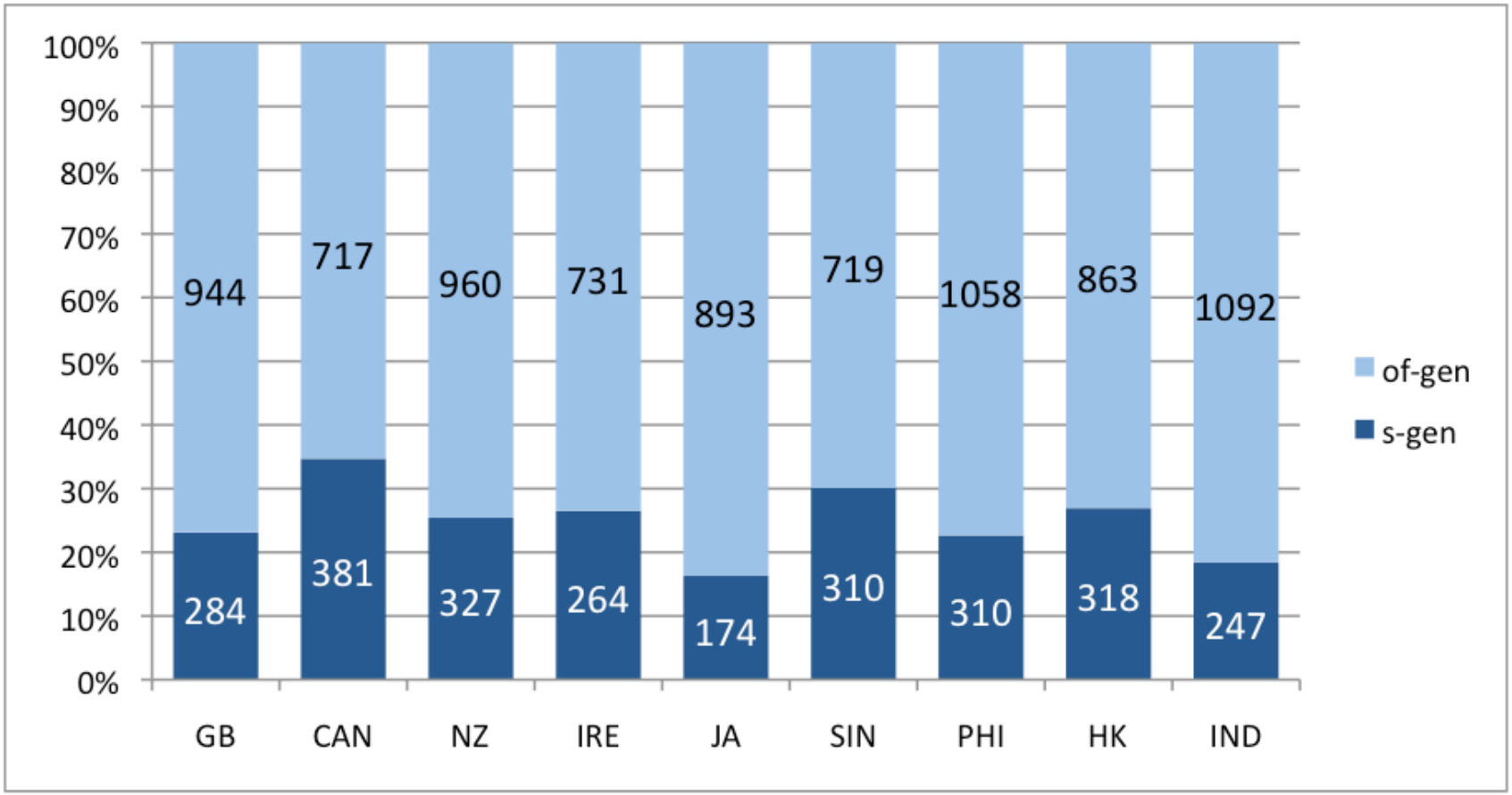
Schneider, E. (2003). The Dynamics of New Englishes: From Identity Construction to Dialect Birth. *Language 79*(2), 233–281.

Schneider, E. (2007). *Postcolonial English: Varieties Around the World*. Cambridge University Press.

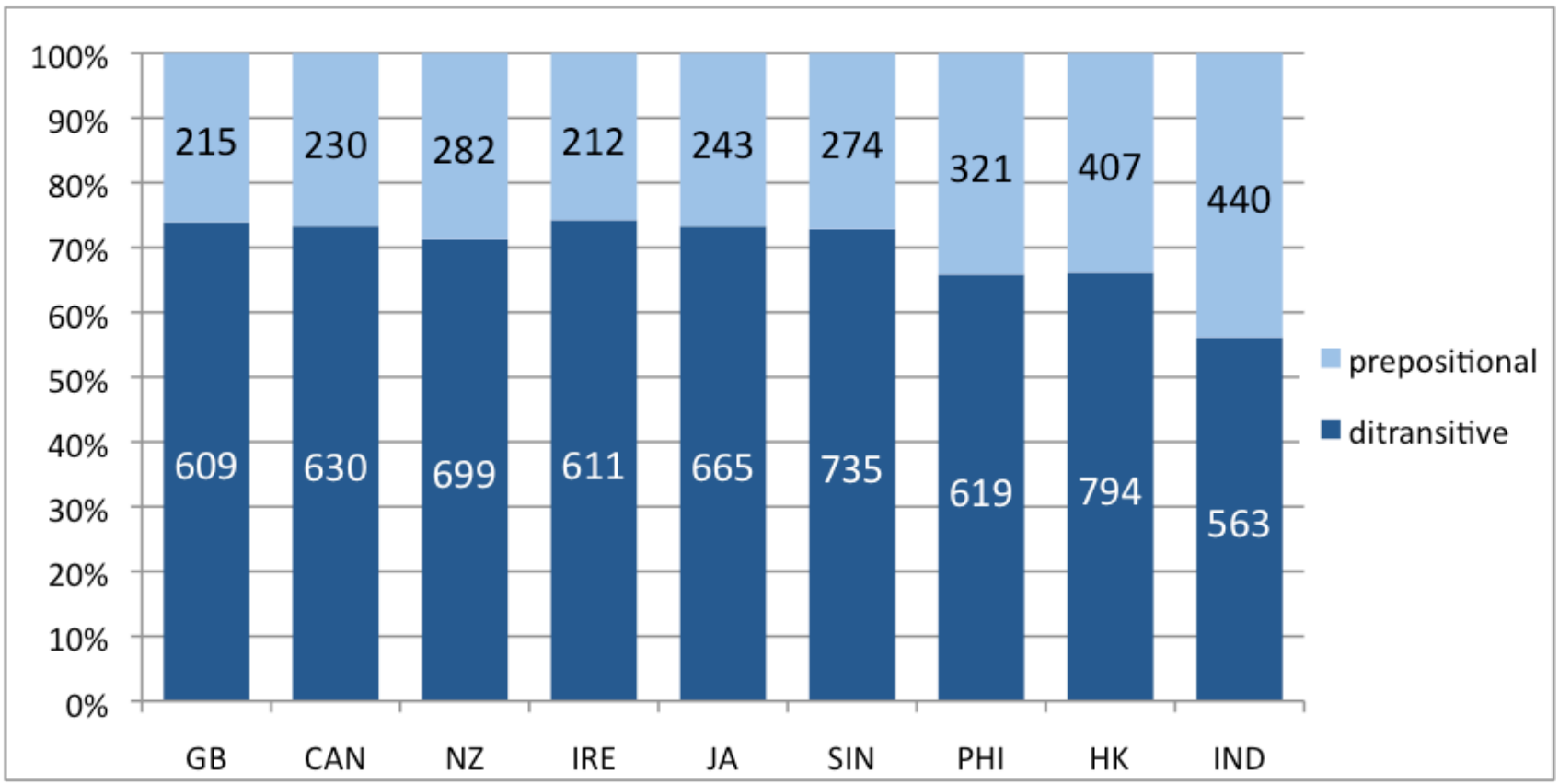
Szmrecsanyi, B., J. Grafmiller, B. Heller, and M. Röthlisberger (2016). Around the world in three alternations: modeling syntactic variation in global varieties of English. *English World-Wide 37*(2).

Wolk, C., J. Bresnan, A. Rosenbach, and B. Szmrecsanyi (2013, January). Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica 30*(3), 382–419.

Proportion of genitives according to variety

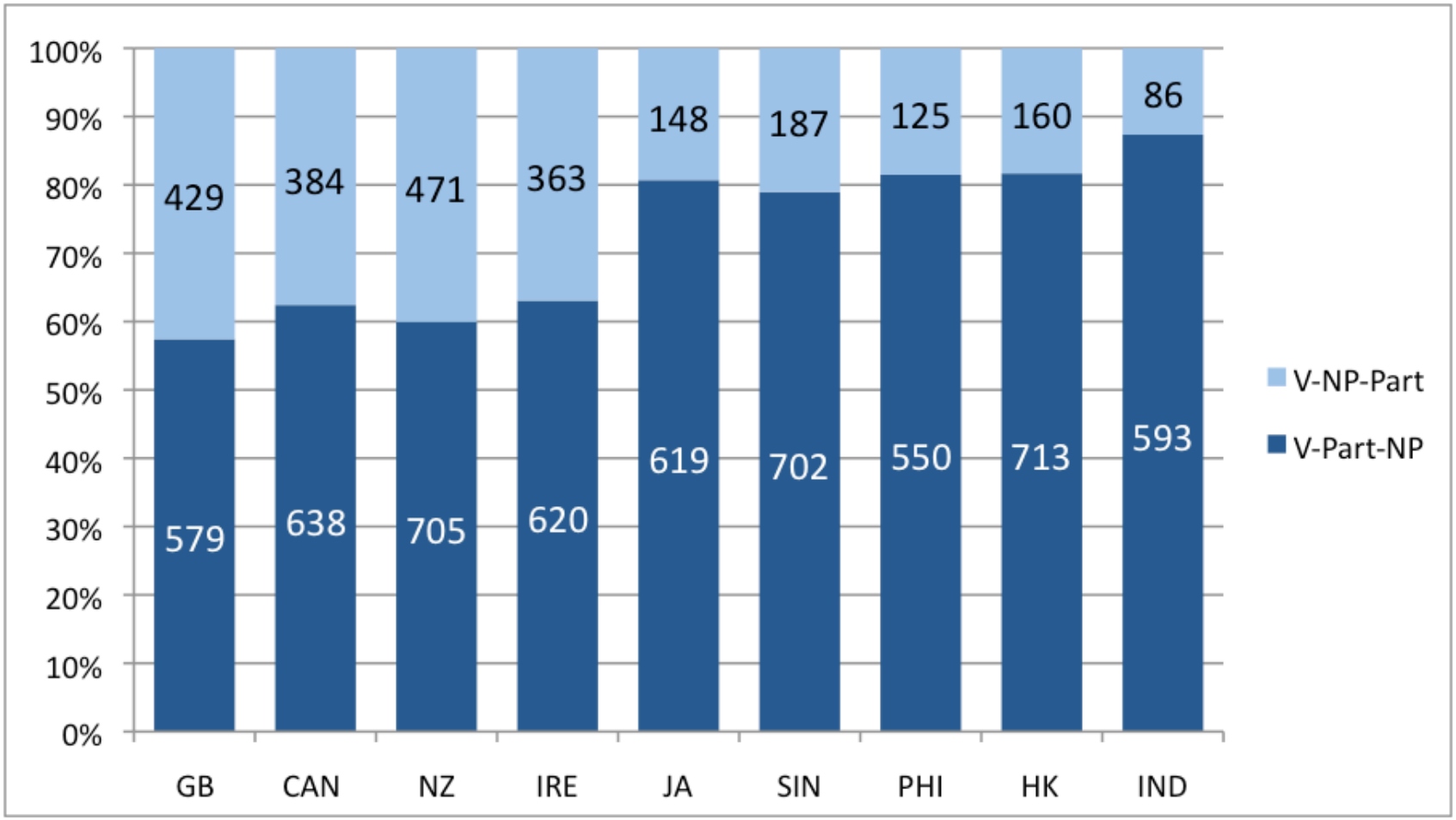


Proportion of datives according to variety



N=8,549

Proportion of particle verbs according to variety



N=8,072